

融合词嵌入表示特征的实体关系抽取方法研究*

张　琴[1,2]　郭红梅[1]　张智雄[1,3]
[1](中国科学院文献情报中心　北京 100190)
[2](中国科学院大学　北京 100049)
[3](中国科学院武汉文献情报中心　武汉 430071)

摘要:【目的】为解决已有方法中单词特征表示不具有语义信息这一问题，对词嵌入表示特征在关系抽取中的作用进行探讨。【方法】考虑词嵌入表示级别、词汇级别和语法级别三种类型特征，利用朴素贝叶斯模型、决策树模型和随机森林模型进行对比实验，并选出代表全部特征的有效特征子集。【结果】使用全部特征时，决策树算法的准确率达到0.48，关系抽取效果最佳，Member-Collection(E_2, E_1)类型关系的F_1值达到0.70，特征排序结果表明依存关系有助于关系抽取。【局限】对小样本量和情况复杂的关系类型识别效果有待提高，以及词向量训练及方法的相关参数需要进一步优化。【结论】实验证明选取的三种类型特征的有效性，词嵌入表示级别特征在实体关系抽取问题中可以发挥重要作用。

关键词: 关系抽取　词嵌入表示　Word2Vec
分类号: TP393

1 引　言

随着网络技术的发展，非结构化信息的数量不断增多，如此庞大的数字资源给人类学习和工作带来困扰。为了更好地利用这些信息，研究人员利用信息抽取技术，将非结构化信息转化成结构化信息。

信息抽取技术希望计算机能够自动识别并抽取出文本中有价值的信息，它具体分为命名实体识别(Named Entity Recognition)、关系抽取(Relation Extraction)、事件抽取(Event Extraction)、时间信息抽取(Temporal Information Extraction)和指代消解(Coreference Resolution)等研究点。其中，关系抽取是指自动识别两个实体之间属于哪种语义关系，例如"John Smith is the chief scientist of the Hard.com Corporation."中实体"John Smith"和"Hard.com"之间存在"Person-Affiliation"的语义关系。实体间关系抽取是信息结构化的重要环节，是知识图谱构建的关键部分，也是问答系统、自然语言理解应用中至关重要的一步。

传统的基于特征工程的实体关系抽取方法中使用单词、实体类型、依存关系等特征，单词使用字典索引表示，这种表示方法不带有语义信息，无法表达两个语义相近的实体之间的联系。词嵌入表示可以将以往离散的单词语义连续化，如果两个单词语义越接近，那么它们对应的词向量空间距离就越大，词嵌入表示为自然语言处理提供非常有效的工具。为了解决以上问题，本文融合词嵌入表示特征进行实体关系抽取。

通讯作者：张琴，ORCID: 0000-0003-1404-842X，E-mail: qinzhang.zq@foxmail.com。
*本文系 ISTIC-EBSCO 文献大数据发现服务联合实验室基金项目"基于 clique 子团聚类的文本主题识别方法研究"的研究成果之一。

融合词嵌入表示特征的实体关系抽取方法考虑词嵌入表示级别特征、词汇级别特征和语法级别特征三类特征，对基于特征工程的实体关系抽取方法进行改进，通过特征排序和有效特征子集进行实体关系抽取效果研究。

2 实体关系抽取相关研究

实体关系定义为两个实体之间的某种联系，用元组 $R=(e_1,e_2)$ 表示，其中 e_1，e_2 是文档 D 中具有关系 R 的实体，关系抽取就是自动找出该特定语义关系。通常，实体关系抽取任务比较关注人、组织、位置等实体之间的关系，例如人和组织之间的"Person-Affiliation"从属关系、组织和位置之间的"Organization-Position"关系。此外，还包括很多其他类别的关系，例如:

①We poured the milk into the pumpkin mixture.

②The burst has been caused by water hammer pressure.

③This article gives details on 2004 in music in the United Kingdom.

句子①中的实体"milk"和"pumpkin mixture"之间存在语义关系"Entity-Destination"；对于句子②和句子③，"burst"和"pressure"存在"Cause-Effect"语义关系，"article"和"music"存在"Message-Topic"语义关系。

关系抽取的研究方法集中于将判断两个实体之间是否存在某种语义关系看作一个分类问题，在此基础上，实体关系分类研究分为核函数方法、远距离监督方法和特征提取方法。

(1) 核函数可以计算结构之间的相似性，实现关系分类目的，效果比较突出的是字符串核函数[1]、解析树核函数[2]、依存树核函数[3]、最短依存路径核函数[4]和多核融合[5]等。其中，Bunescu 等[1]使用词的稀疏子序列、词性标签、通用词性标签、实体类型和 WordNet 同义词等模式，将三种子核函数联合构成字符串序列核函数，通过将它和支持向量机(Support Vector Machine, SVM)模型结合，找到能将正样本与负样本分开的决策超平面。为了解决传统径向基核函数训练矩阵元素趋近于 0 时不利于分类的问题，郭剑毅等[5]对径向基核函数训练矩阵进行改进，并将改进的径向基核函数融合多项式核函数及卷积树核函数，通过枚举的方式获得复合核函数的最优参数，利用多核融合方法与 SVM 模型结合进行中文领域实体关系抽取。

(2) 远距离监督方法利用自举自动产生标注数据，然后训练各种分类器模型完成关系抽取工作[6]。Mintz 等[7]使用 Freebase 知识库，将其中的关系实例所包含的实体同维基百科文本中的实体对齐，从而产生训练数据，然后使用逻辑回归模型进行关系抽取。Banko 等[8]提出 TextRunner 系统，包括学习机、抽取器和评估三个模块。具体过程是：首先，给定一个小样本集，提取两个实体间的单词数量、停用词数量和实体是否是专有名词等特征后，用这组自动标记的特征向量训练朴素贝叶斯分类器得到学习机。然后，抽取器对整个语料库进行单个传递，以提取所有可能的关系元组，将每个元组发送到分类器中，并标记可信赖关系元组。最后，根据文本冗余的概率模型，为每个保留的元组分配概率。远距离监督方法适用于大规模多领域的网络文本信息抽取，使用该方法产生了一系列原型系统，例如 WOE 系统[9]和 ReVerb 系统[10]等。

(3) 特征提取方法利用文本分析处理得到的特征数据训练不同的分类器，特征主要包括实体、词性标签和语法分析结果等。Kambhatla[11]研究实体、实体类型、依存树和解析树等特征，使用最大熵分类器进行关系抽取。Zhou 等[12]考虑两个实体的首单词和 WordNet 中语义类，训练 SVMLight 分类器，研究如何将各种特征组合起来。高俊平等[13]利用词在句子中的位置、词性标签、实体类别、依存关系和语义角色标签等特征，采用条件随机场(Conditional Random Fields, CRF)模型对句子成分进行序列标注，识别中文维基百科数据中概念间的演化关系。甘丽新等[14]在传统特征基础上进行扩展，利用依存句法分析和词性标注结果得到依存句法关系组合特征和最近句法依赖动词特征，使用 SVM 模型作为分类器进行实验。

以往关系抽取研究中的词汇特征往往使用字典索引或独热(One Hot)模型进行表示，在独热模型中单词对应的向量中只有某一维非零，因此，会面临数据稀疏的问题。此外，无论是字典索引表示方法还是独热模型表示方法，单词表示均不带有语义信息，无法识别语义相近的词汇。2013 年，Mikolov 等[15]提出 Word2Vec 词嵌入表示学习模型，旨在将研究对象的语义信息表示为稠密低维实值向量，并且该向量能够表达两个语义相近的单词之间的联系。词嵌入表示模型可以解决数据稀疏和维数灾难问题，在自然语言处理中有广泛应用。

本文融合词嵌入表示特征进行实体间关系抽取，从数据集中提取词嵌入表示级别、词汇级别和语法级别三类特征，将关系抽取看作分类问题，利用这些特征训练朴素贝叶斯模型、决策树模型和随机森林模型，并使用特征排序算法分析各类特征的性能，最后选择有效特征子集，完成关系抽取任务。

3 融合词嵌入表示特征的实体关系抽取方法

基于特征工程的实体关系抽取方法将实体关系识别看作一个分类问题，即将判断两个实体之间是否存在某种关系看作一个分类问题。由此转化为数学问题：文档 $D = w_1, w_2, \cdots, e_1, \cdots, w_j, \cdots, e_2, \cdots, w_n$ 中 e_1 和 e_2 是两个实体，映射函数 f 为：

$$f_R(T(S)) = \begin{cases} +1 & e_1\text{和}e_2\text{之间有}R\text{关系} \\ -1 & e_1\text{和}e_2\text{之间无}R\text{关系} \end{cases}$$

其中，$T(S)$是从文档 D 中提取的特征，通过映射函数 f 判断句子中的实体是否存在关系。这样，实体关系抽取任务等价于实体关系检测任务。

3.1 词嵌入表示

词嵌入表示旨在将单词的语义信息分布式地表示成稠密低维实值向量，单独考虑向量的某一维都没有明确的含义，但是综合考虑这个向量则能够表达这个单词的语义信息，如果两个单词的语义信息相近，则它们的词嵌入表示向量的相似度就越高，空间距离就越小。词嵌入表示研究主要利用神经网络模型进行实现，比较突出的工作有神经网络语言模型(Neural Network Language Model, NNLM)[16]、循环神经网络语言模型(Recurrent Neural Network based Language Model, RNNLM)[17]。2013 年, Mikolov 等提出 Word2Vec[15] 词嵌入表示学习模型，它又细分为两种：一种是 CBOW 模型，已知单词 w_t 的上下文 $w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2}$，预测单词 w_t；另一种是 Skip-gram 模型，在已知单词 w_t 的前提下，预测其上下文 $w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2}$。Word2Vec 模型将神经网络中非常耗时的非线性隐藏层去除，输入是文档集，输出为文档集中的单词的词嵌入表示向量, Mikolov 等的研究表明该模型的词嵌入表示效果较好，因此本文选择 Word2Vec 模型进行词嵌入表示训练。

3.2 特　征

与以往研究不同，融合词嵌入表示特征的实体关系抽取方法创新性地考虑词嵌入表示级别特征，这是由于基于神经网络的词嵌入表示包含单词的语义信息，可以反映词汇之间的语义相关性，本文探究性地考查这种特征的关系抽取效果。同时，考虑词汇级别特征和语法级别特征，研究这三种特征对关系抽取任务的效果。

(1) 词嵌入表示级别特征

按照实体在句子中的相对位置，将左、右两个实体分别记为 E_1 和 E_2。关系抽取工作首先考虑两个实体本身作为特征，两个实体分别用词嵌入表示方法向量化表示为 WE_1 和 WE_2。然后，因为分布式词向量空间存在平移不变性[15]，即 king 和 queen 的向量差与 man 和 woman 的向量差近似相等，所以，本文中的词嵌入表示级别特征考虑实体 E_1 和实体 E_2 的空间向量差 $WE_{12} = WE_1 - WE_2$。此外，具有相同关系的实体对间的语义相似度可能相同或相近，基于这一想法，词嵌入表示级别特征还包括实体对的欧几里德距离和余弦相似性两个特征。假设实体 E_1 的 n 维词嵌入向量表示为 $WE_1 = \{a_1, a_2, \cdots, a_n\}$，实体 E_2 的 n 维词嵌入向量表示为 $WE_2 = \{b_1, b_2, \cdots, b_n\}$。那么，实体 E_1 和实体 E_2 的词嵌入表示向量空间的欧几里德距离如公式(1)所示。

$$D(E_1, E_2) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \tag{1}$$

两个实体 E_1 和 E_2 的词嵌入表示向量空间的余弦相似性如公式(2)所示。

$$S(E_1, E_2) = \frac{\langle WE_1 \cdot WE_2 \rangle}{|WE_1| \times |WE_2|} \tag{2}$$

(2) 词汇级别特征

为了更清楚地描述词汇级别特征，将其细分为词汇特征、类型特征和数量特征。词汇特征主要考虑单词，根据单词在句子中出现位置的不同，将句中所有单词分为三类：两个实体之间的词，实体 E_1 之前的词和实体 E_2 之后的词。因为实体的首单词通常更重要，所以将它们的首单词和其他单词进行区分，考虑两个实体的首单词作为两个特征，分别为 HE_1 和 HE_2。同时，两个实体之间的单词又分为三部分：首个单词，最后一个单词和之间的其他单词。而针对实体 E_1 之前的词和实体 E_2 之后的词，则分别取实体 E_1 前的第一个和第二个单词，以及实体 E_2 后的第一个和第二个单

词。类型特征指实体类型，可以是 ORGANIZATION、LOCATION、DATE、NUMBER、MONEY、PERSON、TIME、ORDINAL、DURATION、MISC 和 OTHER 这 11 类。此外，数量特征主要统计两个实体之间的单词数量和实体数量。按照上述思路，词汇级别特征名称及其详细描述如表 1 所示。

表 1　词汇级别特征及其描述

特征类别	特征	特征描述
词汇	HE_1	实体 E_1 的首单词
	HE_2	实体 E_2 的首单词
	$BNULL$	当实体间没有单词时，取值为 1，否则为 –1
	BO	当实体间仅有一个单词时，取值为该单词，否则为 –1
	BF	当实体间至少有两个单词时，实体间的第一个单词
	BL	当实体间至少有两个单词时，实体间的最后一个单词
	E_1F	实体 E_1 之前的第一个单词
	E_1S	实体 E_1 之前的第二个单词
	E_2F	实体 E_2 之后的第一个单词
	E_2S	实体 E_2 之后的第二个单词
类型	E_1T	实体 E_1 的类型
	E_2T	实体 E_2 的类型
数量	BE	两个实体之间的实体数量
	BW	两个实体之间的单词数量

(3) 语法级别特征

语法级别特征主要指句子的依存解析树中包含的信息和词性标签信息，句子的依存解析树从其句法解析树中获得，包括实体的依存词和实体与其依存词之间的依存关系等信息。具体而言，实体 E_1 和实体 E_2 的依存词分别记为 DE_1 和 DE_2，实体 E_1 和依存词 DE_1 的依存关系记为 R_1，实体 E_2 和依存词 DE_2 的依存关系记为 R_2。词性特征考虑实体 E_1 和实体 E_2 的词性 POS_1、POS_2，实体 E_1 的依存词 DE_1 的词性 $POSD_1$，以及实体 E_2 的依存词 DE_2 的词性 $POSD_2$。

3.3　融合词嵌入表示特征的实体关系抽取方法

融合词嵌入表示特征的实体关系抽取方法基于上述词嵌入表示级别、词汇级别和语法级别三类特征，共 27 个特征，将实体关系抽取工作看作分类问题进行处理。在词嵌入表示特征抽取过程中，针对 Skip-gram 和 CBOW 两种对数线性模型，由于 Skip-gram 模型在识别单词间的语义关系方面效果更好，因此使用 Skip-gram 模型训练词嵌入表示向量。同时，在分类结果方面，本文区分两个实体的顺序，即区分实体关系的方向，例如“Component-Whole (E_1, E_2)”与“Component-Whole(E_2, E_1)”是两种不同的关系，前者表示实体 E_1 是组件，后者表示实体 E_2 是组件。对于训练数据和测试数据，需要计算并提取上述 27 个特征，并利用训练数据的这些特征训练分类器，然后用测试数据检验分类器的关系抽取效果。

4　实验过程与结果分析

4.1　数据集

实验的主要目的是探究本文提出的融合词嵌入表示特征的实体关系抽取方法的有效性，验证其是否能够准确识别实体关系。实验在 SemEval-2010 第 8 个任务[18]提供的数据集上进行，该数据集共有 10 717 个标注样本，其中训练样本 8 000 个，测试样本 2 717 个。这 10 717 个标注样本共包含 9 种有向关系以及 1 种无向关系，有向关系包括“Component-Whole”、“Member-Collection”、“Entity-Origin”、“Entity-Destination”、“Product-Producer”、“Message-Topic”、“Content-Container”、“Instrument-Agency” 和 “Cause-Effect”，无向关系指 “Other”关系。各种关系类型及其所占比例如表 2 所示。

表 2　SemEval-2010 task8 数据集中关系类型及其比例

序号	关系类型	样本数量			占比
		训练集	测试集	总和	(%)
1	Component-Whole(E_2, E_1)	472	150	622	5.80
2	Component-Whole(E_1, E_2)	470	162	632	5.90
3	Member-Collection(E_2, E_1)	612	201	813	7.59
4	Member-Collection(E_1, E_2)	78	32	110	1.03
5	Entity-Origin(E_1, E_2)	568	211	779	7.27
6	Entity-Origin(E_2, E_1)	148	47	195	1.82
7	Entity-Destination(E_2, E_1)	1	1	2	0.02
8	Entity-Destination(E_1, E_2)	844	291	1 135	10.59
9	Product-Producer(E_1, E_2)	323	108	431	4.02
10	Product-Producer(E_2, E_1)	396	123	519	4.84
11	Message-Topic(E_2, E_1)	144	51	195	1.82
12	Message-Topic(E_1, E_2)	490	210	700	6.53
13	Content-Container(E_2, E_1)	166	39	205	1.91
14	Content-Container(E_1, E_2)	374	153	527	4.92
15	Instrument-Agency(E_1, E_2)	97	22	119	1.11
16	Instrument-Agency(E_2, E_1)	407	134	541	5.05
17	Cause-Effect(E_1, E_2)	344	134	478	4.46
18	Cause-Effect(E_2, E_1)	659	194	853	7.96
19	Other	1 407	454	1 861	17.36

4.2 数据预处理

在进行分类实验之前，需要对数据集进行预处理。数据预处理工作包括去停用词、词嵌入表示处理、依存解析树分析、词性标注和关系类型标签数值化等，其中词嵌入表示处理使用Google的Word2Vec工具[19]，训练Skip-gram词嵌入表示模型，经过多次实验，词嵌入表示向量维度大小为100时关系抽取效果最佳。因此，向量维度设置为100，训练窗口的大小设置为5。依存解析树和词性标注等语法分析工作使用斯坦福大学提供的StanfordNLP[20]进行，最后将关系类型标签以1-19进行数值化。

4.3 结合全部特征的关系抽取实验

使用词嵌入表示级别、词汇级别和语法级别27个特征，共324维特征，利用这三类特征训练朴素贝叶斯模型、决策树模型和随机森林模型三种分类器。基于全部特征的实体关系抽取实验使用Python调用scikit-learn实现，分类器使用默认参数和训练集数据进行训练，并利用测试集数据测试它们在关系抽取任务上的性能，分别计算每个分类器的查准率 P、查全率 R 和 F_1 值，结果如表3所示。不使用分类器的情况下，考虑实体关系方向，一个样本被正确分类的概率是1/19，而三种分类器的查准率均大于这一概率，说明融合词嵌入表示特征的实体关系抽取方法的有效性。从表3可以看出，决策树分类器的关系抽取效果最好，其次是随机森林模型，朴素贝叶斯模型的关系抽取效果最差。

表3　分类器的分类效果

分类器	P	R	F_1
朴素贝叶斯模型	0.21	0.21	0.15
决策树模型	0.48	0.47	0.47
随机森林模型	0.45	0.45	0.44

表4是使用决策树模型得到的19类关系的实验查准率 P、查全率 R 和 F_1 值，其中的关系类型标号与表2中的序号相对应，可以看出决策树模型对“Member-Collection(E_2, E_1)”类型关系的 F_1 值达到0.70，查准率、查全率也分别达到0.67，0.73，因此本文中抽取的27个特征对“Member-Collection(E_2, E_1)”这种关系的效果最好。此外，决策树模型对“Entity-Destination (E_1, E_2)”类型关系的查准率、查全率和 F_1 值分别为0.67，0.65和0.66，而对“Entity-Destination(E_2, E_1)”类型关系的查准率、查全率和 F_1 值为0.00的原因是数据集中训练样本和测试样本太少，不能够全面捕捉该类关系的特征。对于“Other”类型关系而言，虽然数据集中的样本数量达到17.36%，但是由于该类型关系情况复杂，所以其的查准率、查全率和 F_1 值不是很高。

表4　各类关系的分类效果

关系类型序号	P	R	F_1
1	0.35	0.30	0.32
2	0.51	0.46	0.49
3	0.67	0.73	0.70
4	0.43	0.31	0.36
5	0.69	0.49	0.57
6	0.38	0.30	0.33
7	0.00	0.00	0.00
8	0.67	0.65	0.66
9	0.42	0.42	0.42
10	0.30	0.30	0.30
11	0.20	0.20	0.20
12	0.39	0.40	0.39
13	0.61	0.64	0.62
14	0.61	0.56	0.58
15	0.07	0.14	0.09
16	0.28	0.30	0.29
17	0.62	0.61	0.61
18	0.61	0.68	0.65
19	0.28	0.31	0.29

4.4 特征排序

本文使用Weka中的ReliefFAttributeEval[21]算法进行特征排序，该算法对特征进行排序的思路是：对于某个特征 a，给出一个样本 A，与样本 A 同类的样本中距离最近的为样本 B，与样本 A 异类的样本中距离最近的为样本 C，评估特征 a 的值时需要考虑样本 B 的特征 a 值和样本 C 的特征 a 值。27种特征排序结果及其所属类型如表5所示。可以看出，前10个特征中有3个是语法级别特征，6个是词汇级别特征，1个是词嵌入级别特征，词汇级别特征信息量更大。其中前3个分别是实体 E_2 的依存词 DE_2、实体 E_1 的首单词 HE_1 和实体 E_2 的首单词 HE_2，这与实体间关系与两个实体本身关系密切相吻合，并且依存关系在实体关系抽取中发挥重要作用。

表 5　特征排序结果

排序	特征	分数	特征类型
1	DE_2	0.0178	语法特征
2	HE_1	0.0152	词汇特征
3	HE_2	0.0104	词汇特征
4	$BNULL$	0.0081	词汇特征
5	R_2	0.0078	语法特征
6	BW	0.0056	词汇特征
7	DE_1	0.0053	语法特征
8	BL	0.0051	词汇特征
9	BF	0.0049	词汇特征
10	WE_1	0.0045	词嵌入特征
11	POS_2	0.0040	语法特征
12	R_1	0.0037	语法特征
13	POS_1	0.0031	语法特征
14	$POSD_2$	0.0031	语法特征
15	$D(E_1, E_2)$	0.0030	词嵌入特征
16	WE_2	0.0027	词嵌入特征
17	$POSD_1$	0.0023	语法特征
18	E_2S	0.0022	词汇特征
19	WE_{12}	0.0015	词嵌入特征
20	E_1F	0.0012	词汇特征
21	E_2F	0.0010	词汇特征
22	E_2T	0.0009	词汇特征
23	E_1T	0.0003	词汇特征
24	BE	0.0002	词汇特征
25	BO	−0.0008	词汇特征
26	$S(E_1, E_2)$	−0.0009	词嵌入特征
27	E_1S	−0.0032	词汇特征

4.5　结合有效特征子集的关系抽取实验

特征选择旨在选择能够代表全部特征的有效特征子集，本文使用 Weka 中的 CfsSubsetEval[22]算法进行特征选择。该算法假设有用的特征子集应该包含那些能够预测分类但彼此间不相关的特征，其构建特征子集的过程是：对于与类别标签相关度最高的特征，只要子集中不包含与它相关度高的特征，则将它添加到特征子集中，迭代处理每一个特征。其优先选择与类别标签相关度高而特征之间相关度低的特征，通过考虑各个特征的分类能力以及特征之间的冗余度，评估特征子集的价值。经过分析得到 $D(E_1,E_2)$ 、 E_1T 、BE、POS_2、$POSD_2$、R_2、$S(E_1,E_2)$ 、BW、$BNULL$、WE_1 、WE_2 、WE_{12} 作为全部特征的特征子集。为了调查该特征子集对关系分类任务的作用效果，使用上述 12 个特征作为全部特征的特征子集，训练朴素贝叶斯模型、决策树模型和随机森林模型三种分类器，实验的查准率 P、查全率 R 和 F_1 值如表 6 所示。

表 6　使用特征子集的关系分类效果

分类器	P	R	F_1
朴素贝叶斯模型	0.16	0.16	0.13
决策树模型	0.44	0.43	0.43
随机森林模型	0.38	0.38	0.37

从表 6 可以看出，对于决策树模型，仅仅使用上述特征子集分类器的 F_1 值也可以达到 0.43，与使用全部特征的 F_1 值相差不大，这说明以上 12 种特征可以作为全部特征的有效特征子集，代表 27 个特征完成实体关系抽取工作。另一方面，上述特征子集中有 5 种是词嵌入表示级别特征，4 种词汇级别特征，3 种语法级别特征，这说明本文提出的词嵌入表示级别特征在关系抽取任务中发挥了重要作用，同时说明本文所选取的三类特征均有效。

5　结　语

本文融合词嵌入表示特征研究实体关系抽取问题，首先将实体用词嵌入方式表示成带有语义信息的低维实值向量，然后从数据集中抽取词嵌入表示级别、词汇级别和语法级别三类特征，最后将实体关系抽取转化为分类问题处理，对比朴素贝叶斯模型、决策树模型和随机森林模型三种分类器的关系抽取效果。实验结果表明综合考虑所有特征时决策树算法的效果最佳，特征排序结果发现词汇级别特征信息量大，依存关系有助于关系抽取，并且利用特征选择算法选择出全部特征的最优特征子集，说明本文选取的三类特征的有效性，且词嵌入表示级别特征在实体关系抽取问题中可以发挥重要作用。

本文的不足之处在于对小样本量的关系类型和语法规则复杂的关系类型存在误判情况。今后的研究将考虑增加上述两种类型的样本数量，同时优化词向量训练的相关参数，从而提高整体识别效果。

参考文献：

[1]　Bunescu R C, Mooney R J. Subsequence Kernels for Relation

Extraction[C]//Proceeding of the 18th International Conference on Neural Information Processing Systems. Cambridge, USA: MIT Press, 2005: 171-178.

[2] Zelenko D, Aone C, Richardella A. Kernel Methods for Relation Extraction[J]. The Journal of Machine Learning Research, 2003, 3(3): 1083-1106.

[3] Culotta A, Sorensen J. Dependency Tree Kernels for Relation Extraction[C]//Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. USA: ACL, 2004: 423-429.

[4] Bunescu R C, Mooney R J. A Shortest Path Dependency Kernel for Relation Extraction [C]// Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. USA: ACL, 2005: 724-731.

[5] 郭剑毅，陈鹏，余正涛，等. 基于多核融合的中文领域实体关系抽取[J]. 中文信息学报，2016，30(1): 24-29. (Guo Jianyi, Chen Peng, Yu Zhengtao, et al. Domain Specific Chinese Semantic Relation Extraction Based on Composite Kernel[J]. Journal of Chinese Information Processing, 2016, 30(1): 24-29.)

[6] Xiang Y, Wang X L, Zhang Y Y, et al. Distant Supervision for Relation Extraction via Group Selection [C]// Proceedings of the 22nd International Conference on Neural Information Processing (Part II). USA: Springer, 2015: 250-258.

[7] Mintz M, Bills S, Snow R, et al. Distant Supervision for Relation Extraction Without Labeled Data [C]// Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. USA: ACL, 2009: 1003-1011.

[8] Banko M, Cafarella M J, Soderland S, et al. Open Information Extraction from the Web [C]// Proceedings of the 20th International Joint Conference on Artificial Intelligence. USA: Morgan Kaufmann Publishers, 2007: 2670-2676.

[9] Wu F, Weld D S. Open Information Extraction Using Wikipedia [C]// Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. USA: ACL, 2010: 118-127.

[10] Fader A, Soderland S, Etzioni O. Identifying Relations for Open Information Extraction [C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. USA: ACL, 2011: 1535-1545.

[11] Kambhatla N. Combining Lexical, Syntactic and Semantic Features with Maximum Entropy Models for Extracting Relations [C]// Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions. USA: ACL, 2004: Article No. 22.

[12] Zhou G D, Su J, Zhang J, et al. Exploring Various Knowledge in Relation Extraction [C]// Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. USA: ACL, 2005: 427-434.

[13] 高俊平，张晖，赵旭剑，等. 面向维基百科的领域知识演化关系抽取[J]. 计算机学报，2016, 39(10): 2088-2101. (Gao Junping, Zhang Hui, Zhao Xujian, et al. Evolutionary Relation Extraction for Domain Knowledge in Wikipedia[J]. Chinese Journal of Computers, 2016, 39(10): 2088-2101.)

[14] 甘丽新，万常选，刘德喜，等. 基于句法语义特征的中文实体关系抽取[J].计算机研究与发展，2016，53(2): 284-302. (Gan Lixin, Wan Changxuan, Liu Dexi, et al. Chinese Named Entity Relation Extraction Based on Syntactic and Semantic Features[J]. Journal of Computer Research and Development, 2016, 53(2): 284-302.

[15] Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and Their Compositionality [J]. Advances in Neural Information Processing Systems, 2013, 26: 3111-3119.

[16] Bengio Y, Ducharme R, Vincent P, et al. A Neural Probabilistic Language Model [J]. Journal of Machine Learning Research, 2003, 3(6): 1137-1155.

[17] Mikolov T, Kombrink S, Burget L. Extensions of Recurrent Neural Network Language Model [C]// Proceedings of 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). USA: IEEE, 2010: 1045-1048.

[18] GitHub [EB/OL]. [2017-05-16]. https://github.com/davidsbatista/Annotated- Semantic-Relationships-Datasets.

[19] Google Code [EB/OL]. [2017-05-16]. http://code.google.com/p/word2vec/.

[20] The Stanford Natural Language Group [EB/OL]. [2017-05-16]. http://nlp. stanford.edu/software/.

[21] Kononenko I. Estimating Attributes: Analysis and Extensions of RELIEF [C]// Proceedings of the European Conference on Machine Learning. USA: Springer, 1994: 171-182.

[22] Hall M A. Correlation-based Feature Subset Selection for Machine Learning [D]. New Zealand: The University of Waikato, 1998.

作者贡献声明:

张琴：提出研究思路，设计研究方案，采集、清洗和分析数据，进行实验，起草论文;

郭红梅：采集、清洗和分析数据，论文修改;

张智雄：论文修改及最终版本修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: qinzhang.zq@foxmail.com。

[1] 张琴. train_test.txt. 实体关系抽取特征集.
[2] 张琴. train.arff. 实体关系抽取特征训练集.
[3] 张琴. test.arff. 实体关系抽取特征测试集.

收稿日期: 2017-06-15
收修改稿日期: 2017-07-12

Extracting Entity Relationship with Word Embedding Representation Features

Zhang Qin[1,2]　Guo Hongmei[1]　Zhang Zhixiong[1,3]
[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)
[3](Wuhan Documentation and Information Center, Chinese Academy of Sciences, Wuhan 430071, China)

Abstract: [**Objective**] This study explores the word embedding representation features for entity relationship extraction, aiming to add semantic message to the existing methods. [**Methods**] First, we used the feature characteristics at word embedding representation, the vocabulary and the grammar levels to extract relations using Naive Bayesian, Decision Tree and Random Forest models. Then, we obtained the optimal subset of the full features. [**Results**] The accuracy of the Decision Tree algorithm was 0.48 with full features, which was the best. The F_1 score of Member-Collection (E_2, E_1) was 0.70, and the dependency could help us extract the relations. [**Limitations**] We need to improve the relation extraction results with small sample size and complex situation. The word vector training method could be further optimized. [**Conclusions**] This study proves the effectiveness of three types of features. And the word embedding representation level feature plays an important role to extract relations.

Keywords: Relation Extraction　Word Embedding Representation　Word2Vec